



A Displacement and Amplitude Score Employing an Optical Flow Technique

CHRISTIAN KEIL AND GEORGE C. CRAIG

Institut für Physik der Atmosphäre, Deutsches Zentrum für Luft- und Raumfahrt, Oberpfaffenhofen, Germany

(Manuscript received 15 December 2008, in final form 9 March 2009)

ABSTRACT

A field verification measure for precipitation forecasts is presented that combines distance and amplitude errors. It is based on an optical flow algorithm that defines a vector field that deforms, or morphs, one image to match another. When the forecast field is morphed to match the observation field, then for any point in the observation field, the magnitude of the displacement vector gives the distance to the corresponding forecast object (if any), while the difference between the observation and the morphed forecast is the amplitude error. Similarly, morphing the observation field onto the forecast field gives displacement and amplitude errors for forecast features. If observed and forecast features are separated by more than a prescribed maximum search distance, they are not matched to each other, but they are considered to be two separate amplitude errors: a missed event and a false alarm. The displacement and amplitude error components are combined to produce a displacement and amplitude score (DAS). The two components are weighted according to the principle that a displacement error equal to the maximum search distance is equivalent to the amplitude error that would be obtained by a forecast and an observed feature that are too far apart to be matched. The new score, DAS, is applied to the idealized and observed test cases of the Spatial Verification Methods Intercomparison Project (ICP) and is found to accurately measure displacement errors and quantify combined displacement and amplitude errors reasonably well, although with some limitations due to the inability of the image matcher to perfectly match complex fields.

1. Introduction

An assessment of the forecast quality of mesoscale numerical weather prediction models is crucial (i) for model development, identifying shortcomings and systematic errors of existing models; (ii) for the documentation of the improvement of forecasting systems in time; and (iii) for the ranking and selection of “good” ensemble members for probabilistic forecasting products and as a key element in novel data assimilation techniques in high-resolution numerical weather forecasting (for more details, see Keil and Craig 2007).

Nowadays, high-resolution numerical models forecast weather with great detail and we might find them useful because observed features are better reproduced. However, the value of these forecasts is difficult to prove using traditional gridpoint-based verification statistics. The classical “double penalty problem” illustrates the limitations of the gridpoint-based error measures: a forecast

of a precipitation feature that is correct in terms of intensity, size, and timing, but incorrect concerning location, results in very poor categorical error scores (many misses and false alarms), and large root-mean-square errors. To address this problem, spatial verification techniques are being developed that do not require the forecasts to exactly match the observations at fine scales. Gilleland et al. (2009, manuscript submitted to *Wea. Forecasting*, hereafter GABCE) classify most of these techniques into one of the four following classes:

- (i) Fuzzy or neighborhood verification techniques require that the forecasts are in approximate agreement with the observations, meaning that forecasts are close in space, time, intensity, or some other important aspect. These techniques typically measure the strength of the agreement as the closeness requirements are varied. Several techniques that have been developed in recent years are summarized in Ebert (2008) and GABCE.
- (ii) Scale-decomposition techniques apply a bandpass spatial filter (e.g., Fourier, wavelet, etc.) so that the scales can be addressed separately. The separation of scales is intended to isolate physical features

Corresponding author address: Dr. Christian Keil, Institut für Physik der Atmosphäre DLR, Oberpfaffenhofen, D-82234 Wessling, Germany.
E-mail: christian.keil@dlr.de

such as large-scale frontal systems or smaller-scale convective showers. An example of this class is the intensity-scale technique (Casati et al. 2004), which measures skill as a function of scale and intensity (e.g., rainfall rates).

- (iii) Feature-based or object-oriented techniques identify weather features (rain systems, cloud features, etc.) in the forecasts and observations and compare their properties. Object-oriented techniques are quite intuitive and effective when the features are well defined and can be associated between the forecast and observations. Examples are the techniques of Ebert and McBride (2000) and Davis et al. (2006).
- (iv) Field verification techniques use optical flow algorithms to compare fields without decomposing them into separate elements or scales. The term optical flow stems from the image-processing community where methods have been developed to represent temporal changes in images as a result of a fluid flowing in a conserved manner. The application of optical flow techniques for forecast verification of cloudiness and precipitation was introduced by Keil and Craig (2007, hereafter KC2007) and Marzban et al. (2008, manuscript submitted to *Wea. Forecasting*).

The purpose of this article is to provide a description of an optical flow based technique, namely the displacement and amplitude score (DAS), and its application to the test cases of the Spatial Verification Methods Intercomparison Project (ICP; Ahijevych et al. 2009, hereafter AGBE).

2. The displacement and amplitude score

The error measure presented in this paper attempts to quantify the difference between a forecast $F(x, y)$ and an observation field $O(x, y)$ in terms of how accurately features are predicted in position and amplitude. The problems of defining what constitutes a feature, and identifying which feature in one image is to be matched with a feature in the other image, are avoided by using an optical flow technique. This method computes a vector field that deforms, or “morphs,” one image into a replica of another, simultaneously displacing all features in the image. The magnitudes of these vectors provide a measure of the displacement error, while the difference between the images after morphing provides a measure of the residual amplitude error.

The optical flow method used here is based on a pyramidal matching algorithm and computes its vector field by seeking to minimize an amplitude-based quan-

tity at successively finer scales within a fixed search environment. The image-matching algorithm and its application to meteorological data are described in detail by Zinner et al. (2008) and KC2007, respectively, and will not be repeated in detail here. An example demonstrating the step-by-step procedure is presented in the next section (in Fig. 2). There are several parameters that must be specified in the pyramidal image-matching algorithm, but as discussed by KC2007, only one has a decisive impact on the resulting vector field. This parameter is the radius of the search environment (maximum search distance), which defines the largest distance over which a feature in one field will be displaced to match a feature in the other field. KC2007 suggest that this should be based on a dynamical scale such as the radius of deformation that characterizes the spatial separation between different synoptic weather conditions. It should be noted that, as with any verification measure, the results will also be influenced by the properties of the fields being matched, such as an intensity threshold for removing background values.

For any feature in the observation field, we can ask how well it is forecast (if at all) in terms of amplitude and location. To do this, the image-matching algorithm is used to deform the forecast field to match the observations. Two fields are constructed: a displacement error field $DIS_{obs}(x, y)$ equal to the magnitude of the displacement vector, and an amplitude error field $AMP_{obs}(x, y)$ defined as the root-mean-square (RMS) difference between the observation field and the morphed forecast field. Both fields are set to zero wherever the observation field is zero, so that errors are only defined where an observed feature is present. A nonzero value of $DIS_{obs}(x, y)$ at the location of an observed feature implies that there was a forecast feature within the maximum search distance, while a zero value means either a perfect location forecast or that no feature was forecast within the maximum search distance. These two possibilities are distinguished by the amplitude error, which will be large for a missed feature.

Similarly, one can ask for each forecast feature how well it corresponds to the observations in amplitude and location. For this, displacement and amplitude error fields for the forecast space error, $DIS_{fct}(x, y)$ and $AMP_{fct}(x, y)$, can be constructed by morphing the observation field onto the forecast field. In this case, a large-amplitude error for a feature where the displacement error is zero indicates a false alarm; that is, something was forecast, but nothing was observed within the maximum search distance. Note that false alarms were not treated correctly by the error measure defined in KC2007, which applied the image matcher only in observation space.

The RMS average of the amplitude errors over the verification domain A gives a scalar amplitude score in observation space:

$$\overline{\text{AMP}}_{\text{obs}} = \frac{1}{n_{\text{obs}}} \left[\sum_A \text{AMP}_{\text{obs}}(x, y)^2 \right]^{1/2}, \quad (1)$$

where n_{obs} is the number of nonzero points in the observation field. A mean forecast space amplitude error $\overline{\text{AMP}}_{\text{fct}}$ is defined analogously. The mean displacement error in observation space is defined as

$$\overline{\text{DIS}}_{\text{obs}} = \frac{1}{n_{\text{obs}}} \sum_A \text{DIS}_{\text{obs}}(x, y), \quad (2)$$

with the mean observation space displacement error $\overline{\text{DIS}}_{\text{fct}}$ defined similarly. The averages are normalized by the number of points where features are found, with n_{obs} and n_{fct} being the number of nonzero pixels in the observation and forecast space, respectively, to make the values insensitive to the size of the domain and the area of nonevents. The total amplitude error (AMP) and displacement error (DIS) are then defined by weighted averages:

$$\text{AMP} = \frac{1}{(n_{\text{obs}} + n_{\text{fct}})} (n_{\text{obs}} \overline{\text{AMP}}_{\text{obs}} + n_{\text{fct}} \overline{\text{AMP}}_{\text{fct}}) \quad (3)$$

and

$$\text{DIS} = \frac{1}{(n_{\text{obs}} + n_{\text{fct}})} (n_{\text{obs}} \overline{\text{DIS}}_{\text{obs}} + n_{\text{fct}} \overline{\text{DIS}}_{\text{fct}}). \quad (4)$$

For many applications, it is not sufficient to have separate amplitude and displacement errors; a single measure of forecast quality is required. Before combining the two components, the displacement error field is normalized by the maximum search distance, D_{max} ; while the amplitude error field is normalized by a characteristic intensity, I_0 , chosen to be typical of the amplitude of the observed features. Analogously to the computation of the amplitude error, the characteristic intensity I_0 is chosen to be the RMS amplitude of the observed field. However, the choice of I_0 depends on the application. For comparing forecast quality over large datasets, the characteristic intensity I_0 could be specified by a climatological rain rate, for instance.

The normalization is based on the principle that, for an observed feature with the characteristic amplitude, a forecast displaced by the distance D_{max} gives the same error as a miss plus a false alarm; that is, no forecast feature is found within a distance D_{max} , but rather there are two unrelated errors in widely separated regions.

The final displacement-amplitude score, DAS, is defined as the average of the two normalized components:

$$\text{DAS} = \frac{\text{DIS}}{D_{\text{max}}} + \frac{\text{AMP}}{I_0}. \quad (5)$$

The DAS values are bounded from below by zero (for a perfect forecast), and will typically take values of order one, although there is no upper limit. A value of one would result from a forecast with the correct amplitude, but an average position error of D_{max} , or a forecast with the correct position with an RMS amplitude error of I_0 , or a combination of both types of error, as will be shown for the real cases in the next section.

3. DAS performance for ICP cases

In the ICP of spatial verification measures there are three different sets of test cases (information online at <http://www.ral.ucar.edu/projects/icp/>) on which DAS has been applied. Selected cases will be presented here in detail to illustrate various properties of the DAS measure. These calculations use a maximum search distance D_{max} of 360 km, corresponding to 90 points at 4-km resolution. For the precipitation fields, an intensity threshold of 1 mm was applied, and a characteristic amplitude I_0 was determined by the RMS average of all observed precipitation values that exceeded the threshold.

a. Geometric cases

The geometric cases are characterized by elliptical precipitation features (axes of the observed feature measure 50 and 200 points) having two different intensities that are designed to help diagnose typical model deficiencies like displacement, aspect ratio, and bias errors (AGBE).

First, the behavior of the displacement and amplitude error fields in observation space is presented for geometric case 1 (pure displacement of the forecast feature by 50 points without any overlapping of both features). The observation is shown in Fig. 1a and the misplaced forecast superimposed with the displacement vector field necessary to minimize the difference between both images in Fig. 1b. Comparison of the morphed forecast (Fig. 1c), in which the displacement vector field is applied on the forecast, with the original observation (Fig. 1a) illustrates an almost perfect match. The magnitude of the displacement vector field within the observed features boundary [only those are considered in $\text{DIS}_{\text{obs}}(x, y)$] is fairly uniform (Fig. 1d), while the amplitude error $\text{AMP}_{\text{obs}}(x, y)$ of the observed and morphed forecast

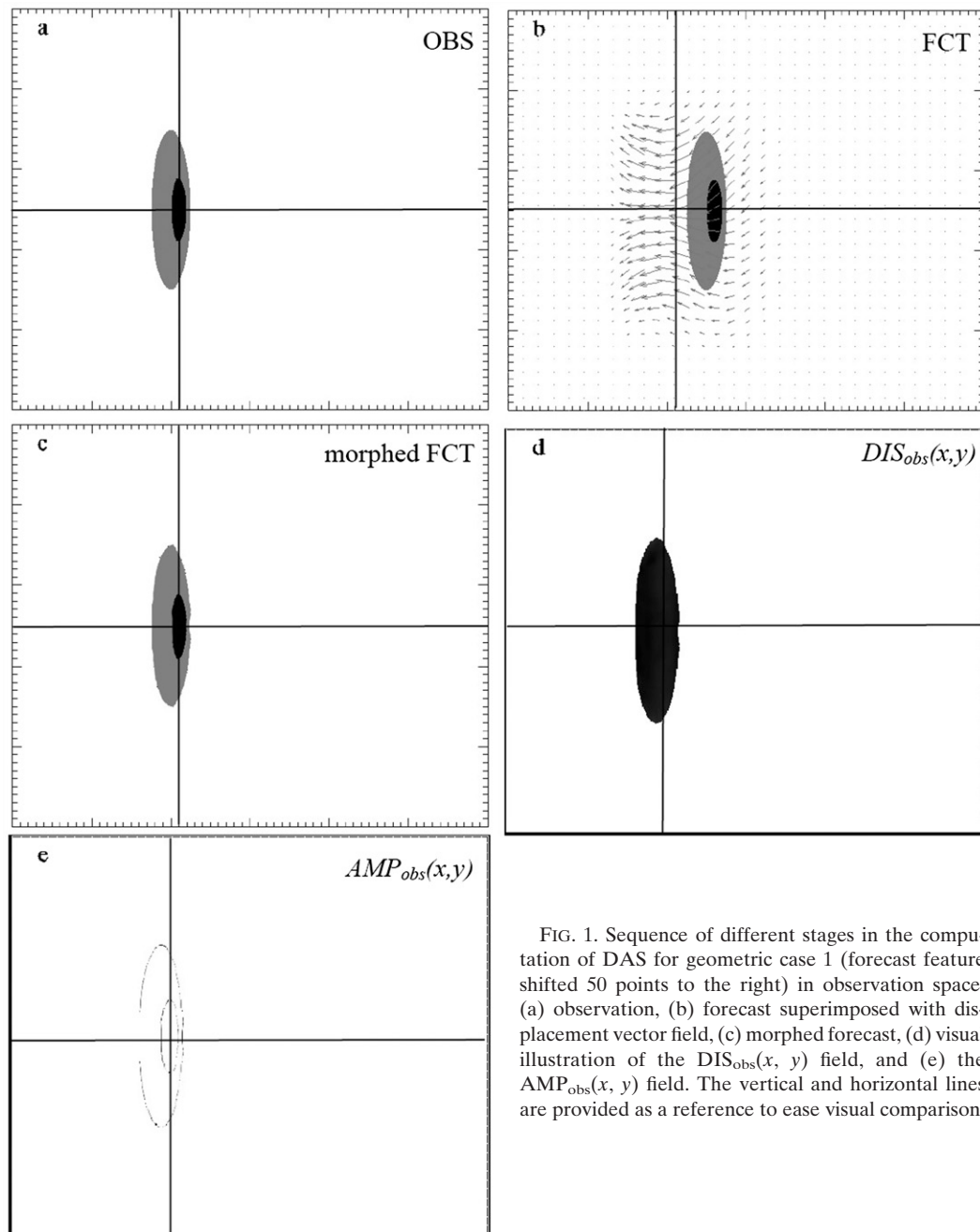


FIG. 1. Sequence of different stages in the computation of DAS for geometric case 1 (forecast feature shifted 50 points to the right) in observation space: (a) observation, (b) forecast superimposed with displacement vector field, (c) morphed forecast, (d) visual illustration of the $DIS_{obs}(x, y)$ field, and (e) the $AMP_{obs}(x, y)$ field. The vertical and horizontal lines are provided as a reference to ease visual comparison.

fields given by the RMS shows small residual errors at the feature boundaries (Fig. 1e), a consequence of interpolation during morphing.

An example sequence illustrating the functioning of the pyramidal image-matching algorithm is presented in Fig. 2. The first three panels (Figs. 2a–c) display the fields at lowest resolution. At this coarsest grain, 16×16 points are averaged to one pixel element. The next three panels (Figs. 2d–f) depict fields at the next higher resolution, where 8×8 points are averaged. Convergence

in the vector field at the coarsest resolution (Fig. 2c) shrinks the morphed feature (Fig. 2e), but this is corrected by the divergent contribution at the next finer resolution (Fig. 2f). At the next higher resolution, the observation (Fig. 2g) and morphed forecast (Fig. 2h) fields are broadly similar, and the vector field at this scale only acts locally at the feature boundaries. The highest resolution ($F = 1$) is not shown. Summing the vector fields over all averaging levels gives the final displacement vector field shown in Fig. 1b.

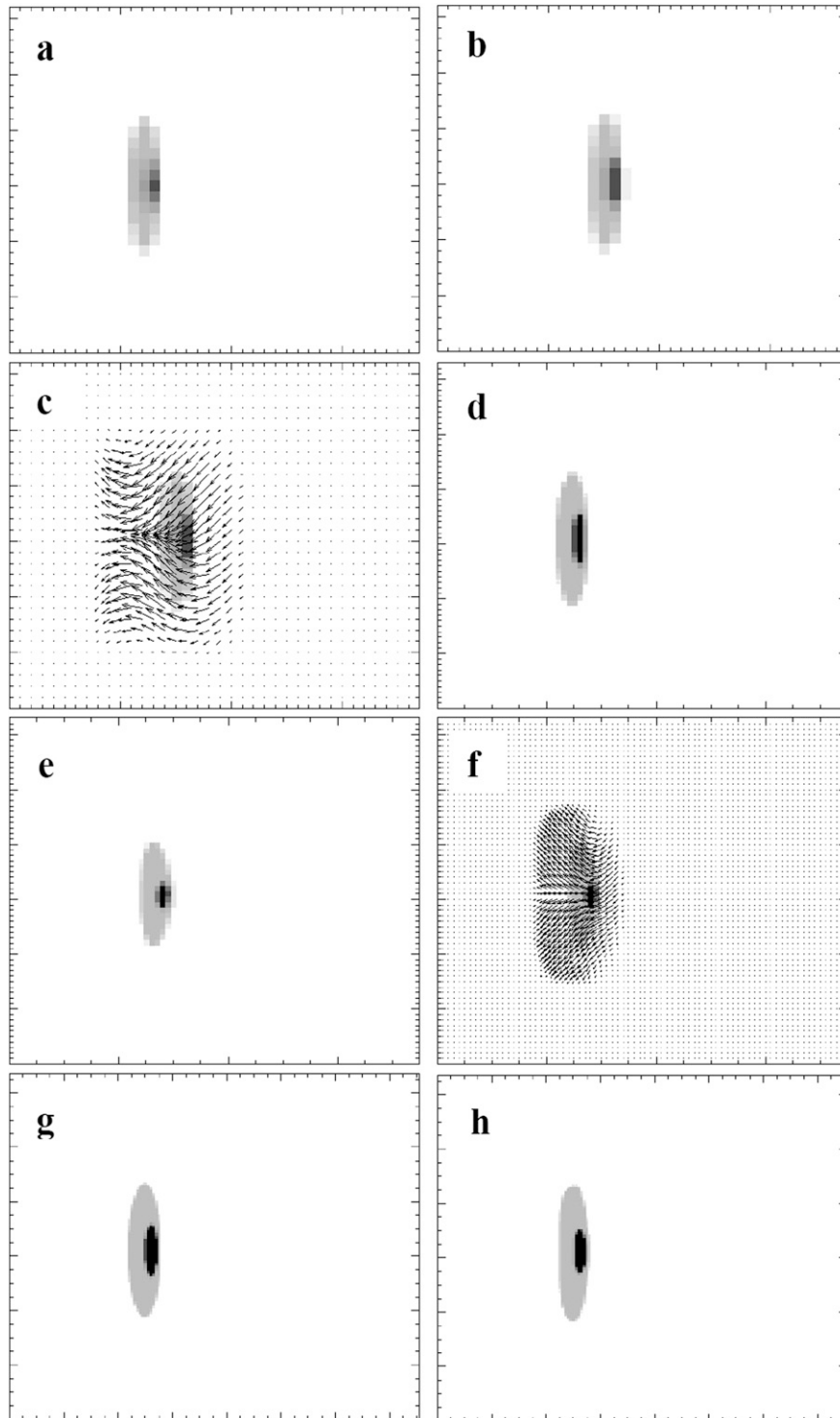


FIG. 2. Sequence of differently coarse-grained fields for geometric case 1: (a) observed and (b) forecast field at lowest resolution where one pixel element contains 16×16 points. (c) The forecast field with the displacement vector field morphing (b) onto (a). (d)–(f) The forecast, the morphed forecast field [after applying the displacement vector field in (c)], and the displacement vector field at the next higher resolution (8×8 pixels compose one pixel element), respectively. (g),(h) Finally, the forecast and the morphed forecast fields at a resolution with 4×4 points composing one pixel element.

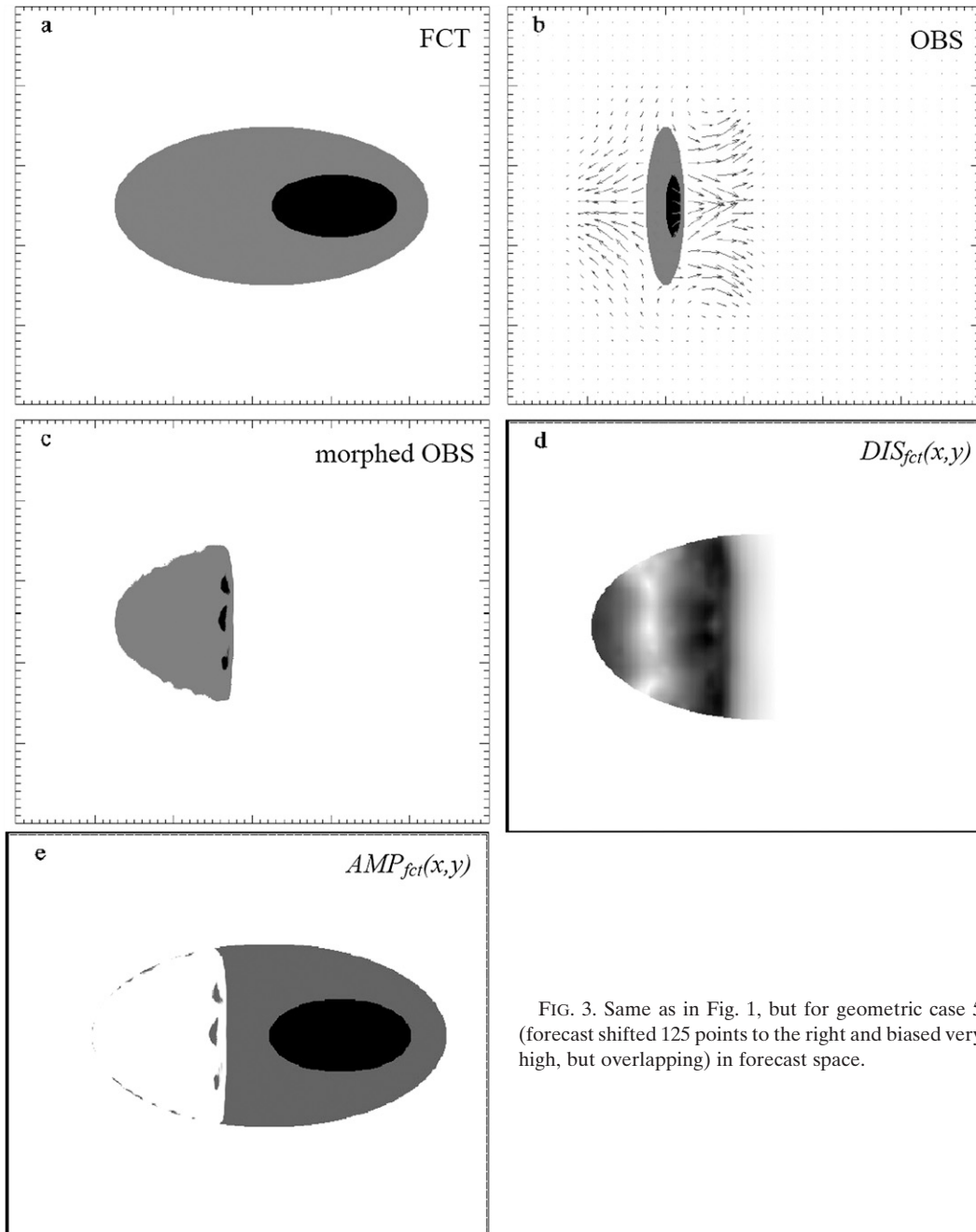


FIG. 3. Same as in Fig. 1, but for geometric case 5 (forecast shifted 125 points to the right and biased very high, but overlapping) in forecast space.

For geometric case 5 (forecast feature much larger in size, displaced but still overlapping the observed feature), the corresponding sequence of images is shown in forecast space (Fig. 3) and observation space (Fig. 4). In forecast space, the overestimated size of the forecast feature (Fig. 3a) results in a strongly divergent displacement vector field, so that the morphed observation field matches the left part of the huge ellipse seen in the forecast (Fig. 3c). The ability of the image-matching

algorithm to stretch the observed field is limited by the specified maximum search distance and thus only part of the forecast feature is regarded as displaced, while the rest is regarded as a forecast “false alarm.” This is clearly seen in the components $DIS_{fct}(x, y)$ and $AMP_{fct}(x, y)$ (Figs. 3d and 3e respectively). In contrast, in observation space a convergent vector field is generated (Fig. 4b), morphing the left side of the forecast feature to match the observations, and shrinking the

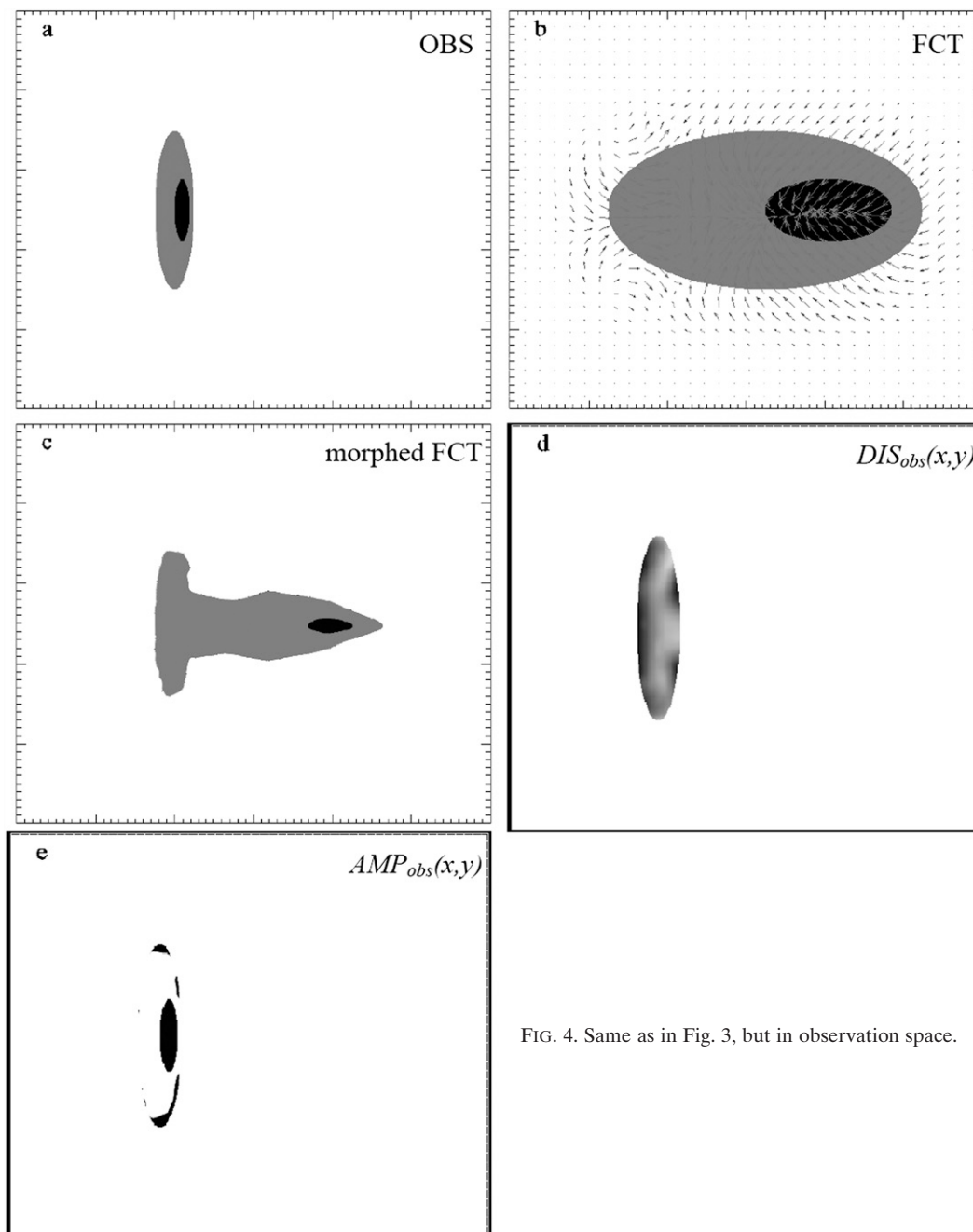


FIG. 4. Same as in Fig. 3, but in observation space.

remaining part (Fig. 4c). Again due to the limitation of the maximum search distance, the excess area of the forecast feature is not completely removed, but this does not contribute to the error amplitude of the observed feature $AMP_{obs}(x, y)$ but represents a false alarm and is accounted for in forecast space (Fig. 3e). The amplitude error (Fig. 4e) in observation space is mainly due to the region of high intensities, which was too far away in the forecast to match to the observations.

The DAS values listed in Table 1 provide an objective ranking of the forecast quality. Also listed are the normalized displacement and amplitude components, DIS/D_{max} and AMP/I_0 , which show the contribution of each component to the final DAS. In geometric case 1 the forecast feature is displaced by 50 points to the right, which corresponds to 55% of the maximum search distance. This is accurately captured by the DIS component. The small residual AMP error is caused by interpolation

TABLE 1. Summary of geometric cases 1–5 depicting a brief description, the DAS, normalized DIS and AMP values (i.e., DIS/D_{\max} and AMP/I_0 with $D_{\max} = 360$ km and $I_0 = 15.4$ mm), and the corresponding rank.

Case	Description of forecast feature	DAS	DIS/D_{\max}	AMP/I_0	Rank
1	50-point displacement	0.62	0.55	0.07	1
2	200-point displacement	1.00	0.00	1.00	2
3	125-point displacement, biased high	1.11	0.21	0.91	5
4	125-point displacement, wrong aspect ratio	1.09	0.22	0.87	4
5	125-point displacement, biased very high, but overlapping	1.02	0.19	0.83	3

errors during morphing. In contrast, the large forecast feature displacement of 200 points in geometric case 2 is beyond the maximum search distance; thus, no matching is possible and $DIS = 0$, while AMP equals 1 as expected for a false alarm plus a miss. For the other geometric cases, which are mixtures of displacement, bias, and aspect ratio errors, both DIS and AMP make significant contributions, although in all cases the amplitude term AMP is larger, indicating large false alarms. The ranking of the geometric cases using DAS gives reasonable results, agreeing with human expectations. Geometric case 1 scores best since the pure displacement within the maximum search distance is captured by the morphing process. Case 2 has the second best score, with the large displacement of the identical feature detected by the algorithm by a pure AMP error. The forecast of case 5 hugely overestimates the observation, but since there is an overlap, it receives some credit and ranks third. Next in rank is case 4 with the wrong aspect ratio. Geometric case 3 scores worst. Note that traditional scores based on contingency tables indicate no skill (for cases 1–4) and rank case 5 as best because of the overlap (AGBE). This illustrates the potential for some of the traditional metrics to be misleading.

b. Perturbed cases

The perturbed cases are constructed using the stage II radar rainfall analysis at 0000 UTC 1 June 2005 as the “observation” and increasingly displacing the precipitation field to the southeast as the “forecast.” Cases 1–5 are characterized by sequentially doubling the separation. For cases 1–4 this is well captured by the DIS error,

which accurately reproduces the displacement distance in each case (see Table 2). For instance, for case 3, DIS is 0.26, corresponding to a displacement of 26% of the maximum search distance. The AMP error is small and remains fairly constant, representing the limit of the accuracy of the morphing process. This residual amplitude error is somewhat larger than for the geometric case considered in the previous section, presumably because of the greater complexity of the field being matched. For case 5, the displacement exceeds the maximum search distance. A few precipitation features are still matched, though not to their counterparts in the displaced field, while most others are not. Consequently, the AMP error (0.9) dominates the total DAS value. In cases 6 and 7, the precipitation features are displaced by the same magnitude as for case 3, but the intensity is increased by 50% in case 6 and decreased by a small constant amount in case 7. The DIS errors of cases 6 and 7 compare well with case 3, correctly measuring the imposed displacement error. For case 6 the AMP error is, as expected, considerably higher (0.44 versus 0.12 for case 3). Interestingly, the final DAS values for cases 4 and 6 are similar. Ideally, one would expect case 4 (50% of maximum displacement error) to be better than case 6 (25% of maximum displacement and 50% intensity error). In case 6, the amplitude error is less than the expected 50%, since the optical flow algorithm distorts the precipitation field to match areas of similar magnitude, rather than just displacing without modifying the structure. Together with the nonzero residual amplitude error in case 4, the contrast in amplitude error between cases 6 and 4 is only about half of

TABLE 2. Same as in Table 1 but for the seven perturbed cases using $D_{\max} = 360$ km and $I_0 = 8.0$ mm.

Case	Description of perturbation	DAS	DIS/D_{\max}	AMP/I_0	Rank
1	Shift 3 points right, –5 points up	0.18	0.07	0.12	1
2	Shift 6 points right, –10 points up	0.24	0.13	0.11	2
3	Shift 12 points right, –20 points up	0.38	0.26	0.12	3
4	Shift 24 points right, –40 points up	0.69	0.50	0.19	5
5	Shift 48 points right, –80 points up	1.20	0.30	0.90	7
6	Shift 12 points right, –20 points up, and intensity $\times 1.5$	0.69	0.25	0.44	5
7	Shift 12 points right, –20 points up, and intensity -0.01	0.46	0.27	0.19	4

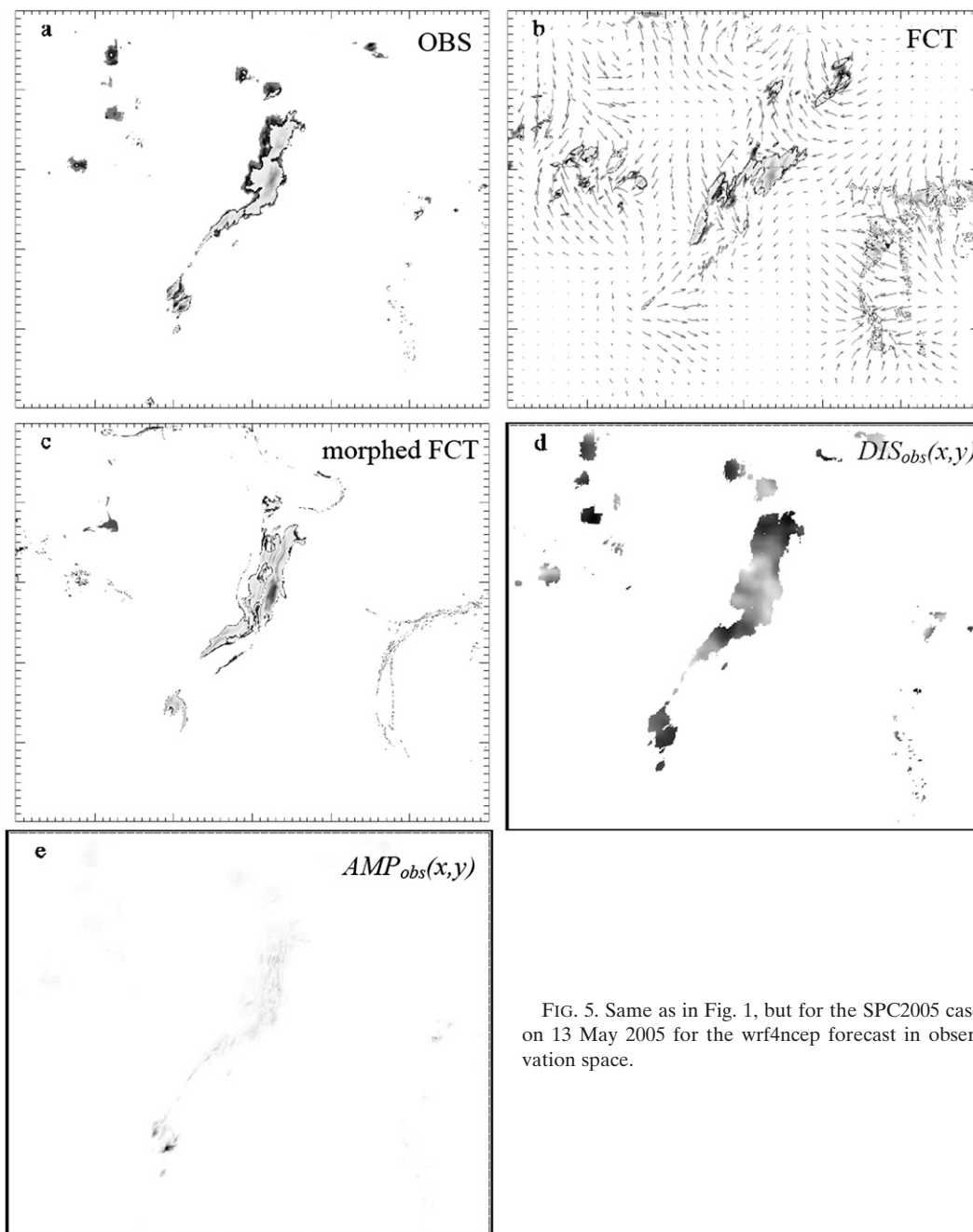


FIG. 5. Same as in Fig. 1, but for the SPC2005 case on 13 May 2005 for the wrf4ncep forecast in observation space.

the expected 50%. It is worth noting that the ranking of the conventional equitable threat score (ETS) is even more counterintuitive, scoring case 6 (0.18) as substantially better than case 4 (0.08).

Summarizing and ranking the DAS values for the seven perturbed cases gives results that agree well with our expectations (Table 2). The perturbed case 1 scores best, since the feature is separated by the smallest distance. Case 1 is followed by cases 2 and 3, in which the features are increasingly separated but within the search

environment. Next in the ranking is case 7, which has the same displacement as case 3 but marginally altered intensities. Case 4, with twice the displacement of case 3, but still within the maximum search distance, follows in the ranking. As previously discussed, case 6 with a large intensity error but medium displacement is tied with case 4. Finally, case 5 is the worst since most precipitation structures are so widely separated in the forecast and observation that they are interpreted as independent errors.

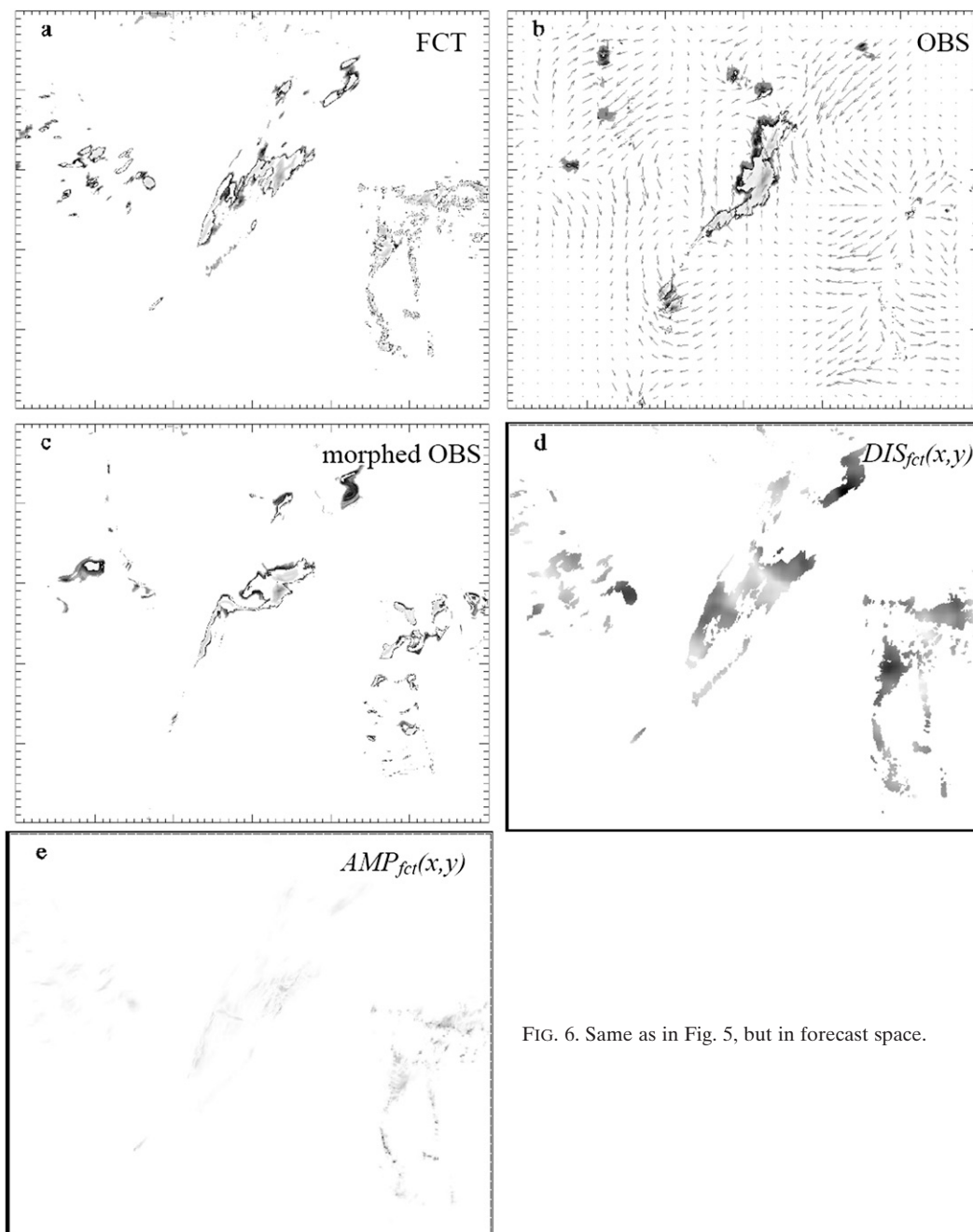


FIG. 6. Same as in Fig. 5, but in forecast space.

c. Real cases

The ICP includes nine 24-h forecasts of 60-min precipitation from each of three different configurations of the Weather Research and Forecasting (WRF) model produced as part of the 2005 Storm Prediction Center's (SPC's) Spring Program (SPC2005; Kain et al. 2008). The performance of DAS will now be discussed in detail for the wrf4ncep forecast on 13 May. This case was chosen to illustrate points made during the discussion at the ICP

workshop (AGBE) of how to include false alarms in the optical flow based error measure. Finally, the DAS results are put in context with a subjective expert ranking and traditional scores for all nine cases in Table 3.

A sequence of images is presented in observation and forecast space in Figs. 5 and 6 (similar to Figs. 3 and 4 for geometric case 5). Comparing the observation in Fig. 5a with the forecast in Fig. 5b, it can be seen that the main differences are that the north–south extension of the main precipitation area (squall line) is underestimated

TABLE 3. Summary of DAS, and normalized DIS and AMP (using $D_{\max} = 360$ km and $I_0 = 6.23$ mm) values with corresponding ranking of the three high-resolution models for all nine cases from SPC2005. Additionally, the values of the subjective evaluations of 24 experts who were asked to rate the forecasts on a scale from 1 to 5 with 1 being poor and 5 being excellent; and the two traditional scores, BIAS and ETS are given for completeness.

Date	Case/model	DAS	DIS/ D_{\max}	AMP/ I_0	Rank	Expert score	BIAS	ETS
26 Apr	wrf2caps	0.81	0.28	0.53	1	3.19	1.46	0.05
	wrf4ncar	0.83	0.32	0.51	3	3.22	1.43	0.06
	wrf4ncep	0.81	0.30	0.51	2	3.40	1.47	0.04
13 May	wrf2caps	1.18	0.30	0.88	2	3.62	1.10	0.17
	wrf4ncar	1.12	0.22	0.90	1	3.61	0.95	0.24
	wrf4ncep	1.38	0.29	1.09	3	2.35	1.45	0.10
14 May	wrf2caps	0.99	0.23	0.75	1	2.62	1.17	0.13
	wrf4ncar	1.08	0.23	0.85	2	2.78	1.06	0.13
	wrf4ncep	1.30	0.28	1.02	3	2.28	1.26	0.07
18 May	wrf2caps	1.09	0.32	0.77	3	3.19	0.95	0.14
	wrf4ncar	1.05	0.28	0.77	2	3.28	0.99	0.11
	wrf4ncep	1.05	0.25	0.80	1	3.32	1.40	0.08
19 May	wrf2caps	0.97	0.28	0.69	2	2.17	1.74	0.08
	wrf4ncar	1.08	0.27	0.81	3	2.17	1.67	0.08
	wrf4ncep	0.83	0.28	0.55	1	2.93	1.52	0.12
25 May	wrf2caps	1.10	0.33	0.76	1	2.80	0.69	0.05
	wrf4ncar	1.22	0.32	0.90	3	2.58	0.74	0.05
	wrf4ncep	1.11	0.31	0.80	2	2.72	1.06	0.05
1 Jun	wrf2caps	1.28	0.34	0.94	3	3.46	0.91	0.10
	wrf4ncar	1.24	0.28	0.96	2	3.48	0.84	0.10
	wrf4ncep	1.22	0.29	0.93	1	3.03	1.01	0.05
3 Jun	wrf2caps	0.94	0.26	0.69	3	2.89	1.02	0.09
	wrf4ncar	0.94	0.27	0.67	2	2.94	0.85	0.07
	wrf4ncep	0.86	0.23	0.63	1	3.19	1.20	0.10
4 Jun	wrf2caps	1.14	0.29	0.85	2	2.49	1.28	0.05
	wrf4ncar	0.91	0.27	0.64	1	3.02	1.06	0.06
	wrf4ncep	1.26	0.31	0.95	3	2.10	1.99	0.05

(miss) and there is spurious precipitation predicted west of the squall line and, in particular, in the southeast (false alarm). In observation space the pyramidal matching algorithm stretches the main precipitation area from the forecast meridionally and tries to diminish the precipitation area in the southeast with a strongly converging vector field (Figs. 5b and 5c). In the corresponding $\text{AMP}_{\text{obs}}(x, y)$ field conditioned for points exceeding the threshold in the observation, the large regions of false alarm are not represented (Fig. 5e). On the other hand, in forecast space, there is a strongly diverging vector field in the southeast area since the algorithm is trying to enlarge the observed rainfall area (Figs. 6b and 6c). Meanwhile, the main precipitation area in the center of the domain is shrunk to match the forecast. In the $\text{AMP}_{\text{fct}}(x, y)$ field the area of false alarms in the southeast is clearly visible (Fig. 6e).

The DAS value for wrf4ncep on 13 May amounts to 1.38, resulting from a large contribution of the AMP error caused by the small-scale high-intensity feature in the south (Fig. 5e) and the false alarms in the southeast (Fig. 6e). Consequently, this forecast is ranked as worse

than the other two model forecasts, in agreement with the subjective ranking of 24 experts (Table 3). For the 13 May case, the bias score (BIAS) and equitable threat score (ETS) (Ebert et al. 2003) confirm the DAS results ranking wrf4ncep as being the worst forecast at this time (BIAS = 1.45 and ETS = 0.10).

Finally, comparison of the DAS values for all nine SPC2005 cases show that a clear ranking of the three models concerning quantitative precipitation forecast quality is not possible. On average, wrf4ncar performs slightly better than wrf2caps and wrf4ncep, consistent with the subjective evaluation (Table 4), but the difference is not large in either ranking. All models perform best on 26 April, where moderate precipitation intensities lead to comparably small-amplitude errors. The worst performance is identified for the previously discussed forecast on 13 May. In general, the values of DAS, the human-generated expert score, and the traditional scores do not appear to be particularly well correlated, although this is perhaps not surprising since each score emphasizes different aspects and properties of the precipitation fields.

TABLE 4. Comparison of mean values of DAS and expert scores averaged over all nine SPC2005 cases for all three high-resolution models.

Model	DAS	Expert score
wrf2caps	1.0534	2.94
wrf4ncar	1.0515	3.01
wrf4ncep	1.0905	2.81

4. Discussion

Progress in weather forecast models has lead to substantially improved and more realistic-appearing forecast fields. However, traditional verification measures often indicate poor performance because of the increased small-scale variability. As a result, the true value of high-resolution forecasts is not always characterized well. To address this problem, spatial verification techniques are being developed that do not require the forecasts to exactly match the observations at fine scales. One promising class of spatial verification methods makes use of optical flow techniques in quantifying spatial differences between the forecast and observation fields. The new displacement and amplitude score DAS, proposed in this paper, relies on a computationally efficient pyramidal image-matching algorithm (~ 10 seconds per image pair on a PC). To account for false alarms and misses, the algorithm is applied in observation space, morphing the forecast onto the observation, and in forecast space, morphing the observation onto the forecast. The contributions from observation and forecast space are averaged to give scalar amplitude and displacement scores. For applications that require a single measure of forecast quality, the separate amplitude and displacement errors are combined. To do this, the displacement error field is normalized by the maximum search distance, while the amplitude error field is normalized by a characteristic intensity chosen to be typical of the amplitude of the observed features, and the two normalized errors are summed.

Within the framework of ICP DAS has been applied on all common cases, including the geometric, the perturbed, and the nine cases of SPC2005. The displacement error term accurately measured the distance between the observed and forecast features. In the synthetic experiments with pure displacement errors, the amplitude error term was contaminated by a small residual error, probably a result of interpolation errors in the matching algorithm; however, there was no evidence of such an error in cases where the displacement was combined with an actual amplitude difference. There is however an inherent ambiguity between the displacement and amplitude errors in complex fields, where a forecast object might be regarded as a bad amplitude

forecast of a nearby object, or as a better amplitude forecast of a more distant feature. Matching the forecast and observations through optical flow avoids part of this ambiguity since there is no need to define the criteria used to identify individual objects in the two fields, and is found to provide consistent identification of the displacement and amplitude errors in idealized cases where there is no ambiguity.

Within the context of the ICP, a wide array of possible forecast errors has been addressed. However, closeness in time has not explicitly been considered. Since the errors in high-resolution forecasts are often related to, for example, the mistiming of a frontal passage or the onset of convection, it would be highly desirable to extend the application of DAS to multiple times. The application of DAS to rank and select individual realistic ensemble members to generate probabilistic forecasting products will be explored in future work.

Acknowledgments. We gratefully acknowledge Hermann Mannstein (DLR) for providing the pyramidal matching algorithm.

REFERENCES

- Ahijevych, D., E. Gilleland, B. Brown, and E. Ebert, 2009: Application of spatial verification methods to idealized and NWP gridded precipitation forecasts. *Wea. Forecasting*, in press.
- Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl.*, **11**, 141–154.
- Davis, C. D., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784.
- Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, doi:10.1002/met.25.
- , and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.*, **239**, 179–202.
- , U. Damrath, W. Wergen, and M. E. Baldwin, 2003: The WGENE assessment of short-term quantitative precipitation forecasts. *Bull. Amer. Meteor. Soc.*, **84**, 481–492.
- Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430.
- Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952.
- Keil, C., and G. C. Craig, 2007: A displacement-based error measure applied in a regional ensemble forecasting system. *Mon. Wea. Rev.*, **135**, 3248–3259.
- Zinner, T., H. Mannstein, and A. Tafferner, 2008: Cb-TRAM: Tracking and monitoring severe convection from onset over rapid development to mature phase using multi-channel Meteosat-8 SEVIRI data. *Meteor. Atmos. Phys.*, **101**, 191–210, doi:10.1007/s00703-008-0290-y.